Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Finding Useful Solutions in Online Knowledge Communities: A Theory-Driven Design and Multilevel Analysis

Xiaomo Liu, G. Alan Wang, Weiguo Fan, Zhongju Zhang

To cite this article:
Xiaomo Liu, G. Alan Wang, Weiguo Fan, Zhongju Zhang (2020) Finding Useful Solutions in Online Knowledge Communities: A Theory-Driven Design and Multilevel Analysis. Information Systems Research 31(3):731-752. https://doi.org/10.1287/isre.2019.0911

INFORMATION SYSTEMS RESEARCH
Vol. 31, No. 3, September 2020, pp. 731–752
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Finding Useful Solutions in Online Knowledge Communities: A Theory-Driven Design and Multilevel Analysis

Xiaomo Liu,[a] **G. Alan Wang,**[b] **Weiguo Fan,**[c] **Zhongju Zhang**[d]

[a] S&P Global Ratings, New York, New York 10041; [b] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, Blacksburg, Virginia 24061; [c] Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa 52242; [d] W. P. Carey School of Business, Arizona State University, Tempe, Arizona 85287

Contact: xiaomo.liu@spglobal.com, https://orcid.org/0000-0003-4184-4202 (XL); alanwang@vt.edu, https://orcid.org/0000-0002-5026-881X (GAW); weiguo-fan@uiowa.edu, https://orcid.org/0000-0003-1272-5538 (WF); zhongju.zhang@asu.edu, http://orcid.org/0000-0001-9200-2369 (ZZ)

Received: July 28, 2017
Revised: June 8, 2018; April 6, 2019; August 6, 2019
Accepted: September 10, 2019
Published Online in Articles in Advance: May 7, 2020

https://doi.org/10.1287/isre.2019.0911

Abstract. Online communities and social collaborative platforms have become an increasingly popular avenue for knowledge sharing and exchange. In these communities, users often engage in informal conversations responding to questions and answers, and over time, they produce a huge amount of highly unstructured and implicit knowledge. How to effectively manage the knowledge repository and identify useful solutions thus becomes a major challenge. In this study, we propose a novel text analytic framework to extract important features from online forums and apply them to classify the usefulness of a solution. Guided by the design science research paradigm, we utilize a kernel theory of the knowledge adoption model, which captures a rich set of argument quality and source credibility features as the predictors of information usefulness. We test our framework on two large-scale knowledge communities: the Apple Support Community and Oracle Community. Our extensive analysis and performance evaluation illustrate that the proposed framework is both effective and efficient in predicting the usefulness of solutions embedded in the knowledge repository. We highlight the theoretical implications of the study as well as the practical applications of the framework to other domains.

History: Sabyasachi Mitra, Senior Editor; Wendy Duan, Associate Editor.
Funding: W. Fan is partially supported by the National Natural Science Foundation of P.R.C. [Grants 71531013, 71729001].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2019.0911.

Keywords: online knowledge community • information usefulness • argument quality • source credibility • text mining • theory-driven design science

1. Introduction

Over the last two decades, the concept of communities of practice has become increasingly popular. Various organizations have adopted the concept for the purpose of managing knowledge sharing and learning both internally and externally. According to Wenger (1998), communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly. Three elements are crucial to cultivate a community of practice: an identity defined by a shared domain of interest, engagement among members to develop relationships and learn from each other, and a shared repertoire of resources (Wenger 1998). In an attempt to clarify variations in usage and conceptualizations, Cox (2005) provides a comprehensive and comparative review of communities of practice.

Communities of practice can be organized for different reasons by different entities. In this study, we focus on one organizational form of knowledge community, *internet-enabled online knowledge communities* (OKCs). We adopt the framework developed by Lee and Cole (2003) and describe an OKC as a model for community-based “knowledge creation in purposeful, loosely coordinated, [and] distributed systems” (p. 633). Many firms have built OKCs that allow their employees and customers to interact for a variety of purposes, such as soliciting ideas for new product and service development, building a brand, and providing product-related knowledge and support (Huang et al. 2018). In addition, OKCs can be formed as work, profession, or special interest group communities that are not sponsored by firms (de Vries and Kommers 2004). Table 1 summarizes the major categories of OKCs and identifies an example community in each category. In this study, we have a particular interest in studying the communities for providing product-related knowledge and support, as well as those for sharing work, profession, and special interest knowledge. Both of them focus on problem solving, a knowledge activity

Table 1. Major Categories of Online Knowledge Communities

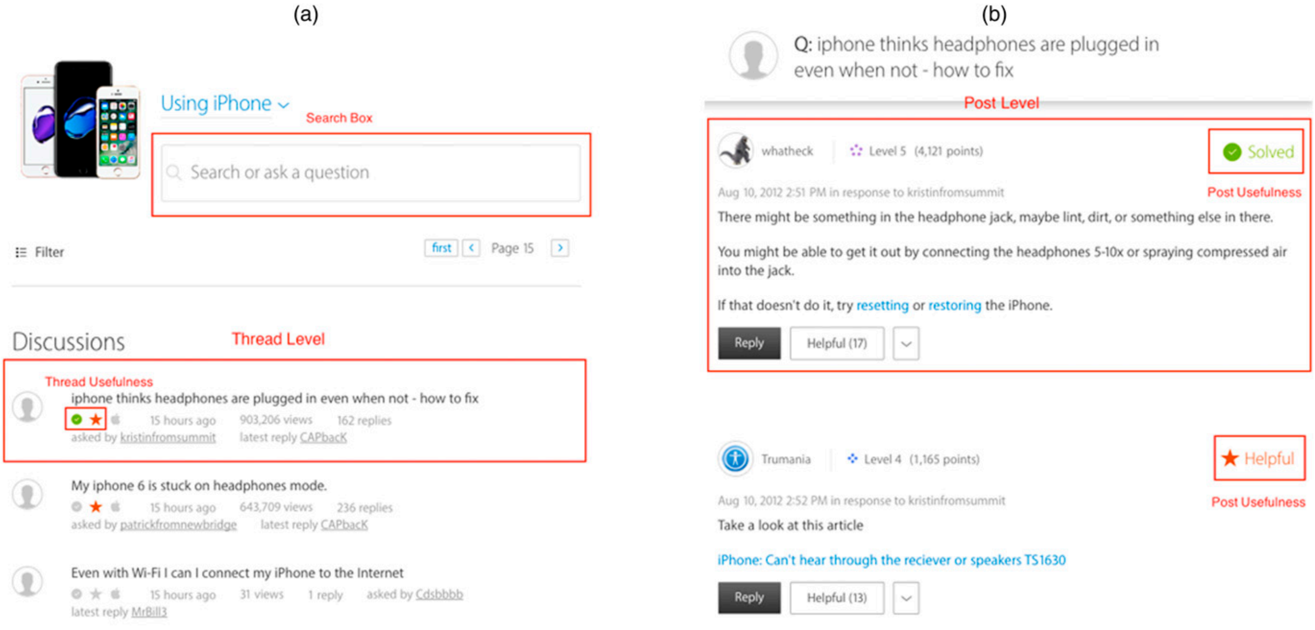
Category	Sponsor	Example community	Problem-solving focused?
Communities for brand building and promotion	Firm	Sephora Beauty Insider Community (http://sephora.com/community)	No
Communities for soliciting ideas for new product and service development	Firm	LEGO Ideas (http://ideas.lego.com)	No
Communities for providing product related knowledge and support	Firm or nonfirm	Apple Support Community (http://discussions.apple.com)	Yes
Communities for sharing work, profession, and special interest knowledge	Firm or nonfirm	Oracle Community (http://community.oracle.com)	Yes

that has long been recognized as a primary vehicle for better understanding of our environment, learning, and discovery of new opportunities (Gray 2001).

OKCs have a number of characteristics that distinguish them from the early development of online communities. First, most knowledge communities have now introduced incentive and reputation mechanisms that motivate members to actively participate in the communities. Second, the questions posted in knowledge communities often come from issues that members encounter in their daily work. Additionally, members with different levels of skills/expertise engage in *informal conversations* responding to questions and answers, and over time, they produce a huge amount of highly unstructured and implicit knowledge.

OKCs usually organize online discussions into two levels: thread-level view and post-level view (Figure 1). In a typical knowledge consumption process, a user starts at the thread view to identify relevant discussions to her questions and then reads through replies at the post view, hoping to obtain a useful solution. Davis (1989) argues that users are more likely to adopt a system for problem solving if it is useful and is easy to use. Therefore, it is valuable that OKCs can offer a mechanism to assist users to find both *relevant* and *useful* information in a timely manner. A common practice is to provide a community-wide search engine. The search feature in OKCs generally works well to find relevant information but often fails to identify useful solutions. Users will still need to read each relevant thread and corresponding posts in the search results. In recent years, some OKCs have introduced a feature allowing the original question seekers to rate and label whether a post is *useful* in answering their questions. However, the rating and

Figure 1. (Color online) An Example of Two-Level Views (Thread (accessed May 23, 2018, https://discussions.apple.com/community/iphone/using_iphone) and Post (accessed May 23, 2018, https://discussions.apple.com/thread/4196806) Levels) of Apple Support in 2018



iphone thinks headphones are plugged in even when not - how to fix
15 hours ago 903,206 views 162 replies
asked by kristinfromsummit latest reply CAPbacK

My iphone 6 is stuck on headphones mode.
15 hours ago 643,709 views 236 replies
asked by patrickfromnewbridge latest reply CAPbacK

Even with Wi-Fi I can I connect my iPhone to the Internet
15 hours ago 31 views 1 reply asked by Cdsbbbb
latest reply MrBill3

Thread View

Q: iphone thinks headphones are plugged in even when not - how to fix

whatheck Level 5 (4,121 points)
Aug 10, 2012 2:51 PM in response to kristinfromsummit
There might be something in the headphone jack, maybe lint, dirt, or something else in there.

You might be able to get it out by connecting the headphones 5-10x or spraying compressed air into the jack.

If that doesn't do it, try resetting or restoring the iPhone.

Trumania Level 4 (1,165 points)
Aug 10, 2012 2:52 PM in response to kristinfromsummit
Take a look at this article

iPhone: Can't hear through the reciever or speakers TS1630

Post View

labeling of post usefulness is voluntary, leading to a systemic "error of omission" problem. Anecdotal evidence from our investigation and other studies such as Cao et al. (2011) show that the percentage of user voting on usefulness is often low. In typical OKCs such as Oracle and Apple Support communities, only about 10% and 27% of threads, respectively, contain usefulness feedback.[1] As a result, there is a strong motivation to identify useful solutions automatically in OKCs to better serve users' knowledge needs and bolster the success of these communities.

From a theoretical perspective, information usefulness has different degrees to which a person perceives information to be valuable, informative, and helpful (Davis 1989, Sussman and Siegal 2003, Sarker and Valacich 2010). As such, original question seekers in OKCs are allowed to rate multiple "helpful" posts and, from them, select one "solved" post that is the best solution. Automatic identification/classification of information usefulness at various levels can be utilized by OKCs in many ways such as building a knowledge base of answered questions, preventing duplicated questions automatically (given that there is a good algorithm to match new and answered questions), or ranking questions with good answers higher in the search results. Therefore, the main research question of this study is how to design a machine learning algorithm that can predict useful solutions in OKCs without human intervention. To do that, we propose a novel text analytic framework to automatically extract a comprehensive set of features, which will then feed into a classification algorithm to predict the probability of usefulness of a solution at multiple levels (post or thread level, useful or solved).[2]

The design of our framework is guided by an adequate process model of information usefulness (Kraaijenbrink et al. 2007). Such a process model provides a foundation and serves as a kernel theory for the design (Walls et al. 1992). In our comprehensive literature search, we found that the majority of existing studies on usefulness mining are primarily computational methods that ignore the implementation of design theories. To address this drawback, we operationalize the knowledge adoption model of information usefulness (Sussman and Siegal 2003) in designing our text analytic framework. From this model, we assess the usefulness of a solution using a comprehensive set of indicators for argument quality and source credibility in OKCs. We test our framework on two large-scale OKCs, the Apple Support Community and Oracle Community, through an iterative search process of the feature sets and a rigorous evaluation of various classification algorithms. Our extensive analysis and performance evaluation illustrate that the proposed framework is both effective and efficient in predicting the usefulness of solutions embedded in the knowledge repository of OKCs. It is worth noting that our work tackles the problem of knowledge management in OKCs from a different angle than do prior information retrieval (IR) studies. It estimates the *usefulness*, which is above and beyond the *relevance* focused on by IR (i.e., a search engine can retrieve an online discussion relevant to the topic of one's question, but it does not necessarily contain answers to resolve that question). Furthermore, our framework utilizes a rich set of domain relevant contextual cues that measure both the source credibility and the content argument quality, which are not the focus of traditional IR algorithms.

The rest of this paper is organized as follows. In Section 2, we review relevant streams of conceptual and computational studies in information usefulness as well as recent advances in the application of the design science methodology. In Section 3, we follow the information system design theory (ISDT) to operationalize the knowledge adoption model and develop various dimensions of information usefulness as well as associated quantitative metrics. Section 4 presents the overall architecture of our framework. After discussing our data and research setting in Section 5, we perform extensive evaluation of our algorithm and compare its performance with various benchmarks in Section 6. Section 7 summarizes and provides implications, concluding remarks, and future research directions.

2. Background and Related Work

Text-based analytic systems are a related class of systems that use either text categorization and analysis (text mining) or IR technologies (Abbasi and Chen 2008a). As pointed out by Hearst (1999), the goal of text mining is to discover new information from text (Fan et al. 2006). The information systems (IS) research has advanced text mining to analyze online user-generated content and derived insights to support various decision-making tasks. CyberGate is an example system that is capable of extracting a comprehensive set of linguistic features to conduct sophisticated text mining (Abbasi and Chen 2008a).

This study follows a similar philosophy and deals with identifying useful solutions to the questions raised in OKCs. The terminology of "usefulness" or "helpfulness" has been used independently in many related studies (Mudambi and Schuff 2010, Racherla and Friske 2012). These two terms are used interchangeably in this research, as they both represent users' satisfaction to their information needs. Prior research on this topic has primarily focused on developing either empirical and conceptual models (from information systems) or computer algorithms and systems (mainly from computer science) to study information usefulness. In conceptual studies, researchers used theories from social sciences to model the factors that can affect perceived usefulness and used hypothesis testing on real data to validate their models (Gregor 2006). These studies provided a thorough understanding of the underlying mechanisms that govern the decision-making process of usefulness. A further question to be answered is how effective the conceptual models are in guiding the development of computational models for real IS systems. In the following, we take a closer look at the related studies of this topic.

2.1. Empirical and Conceptual Models of Information Usefulness

There is a rich literature that studies user-generated content (UGC) (including online reviews, discussion forums, blogs, etc.). In one stream of research, scholars examine the relationship between online reviews and product sales. For instance, Chevalier and Mayzlin (2006), Duan et al. (2008), and Chintagunta et al. (2010) find that improved user sentiment or perception in online consumer reviews (especially user ratings) leads to an increase in the relative sales of products (such as books and movie box office sales). In a separate study by Duan et al. (2008), who consider the endogenous nature of online reviews, the authors find that the rating of online user reviews has no significant impact on movies' box office revenue even though the volume of online reviews does. Forman et al. (2008) show that the prevalence of review disclosure of identity information is associated with an increase in online product sales. The temporal effects of online reviews on sales, such as self-selection biases (Li and Hitt 2008) and diminishing impact (Hu et al. 2008), have also been examined in this stream of research.

Our study is closer to the literature that examines the helpfulness of UGC from different online communities or e-commerce platforms. The majority of this stream of research use either primary survey data or secondary archived online review data to identify the impact of different factors on the perceived helpfulness of online reviews (e.g., Mudambi and Schuff (2010), Cao et al. (2011), Pan and Zhang (2011), Baek et al. (2012), Ngo-Ye and Sinha (2014), Yin et al. (2014), Chua and Banerjee (2015), Huang et al. (2015), Hong et al. (2017), Karimi and Wang (2017), Zhou and Guo (2017), and Malik and Hussain (2018)). Our extensive literature review finds that the determinants of review helpfulness mainly fall into two categories: (1) review-related factors, such as review length, review readability, emotions, review ratings, and review age; and (2) reviewer-related factors that are derived from reviewers' self-disclosure or social actions, including reviewer tenure, expertise, badges, image, centrality, and the number of followers of a reviewer. These two categories of determinants also map well to the knowledge adoption model (see Section 3), which provides us the theoretical foundation for the design of our analytic framework.

Despite the above efforts to understand the causes and effects of online reviews and other knowledge sharing behaviors (Cheung et al. 2013, Majchrzak et al. 2013, Haas et al. 2015), little IS research has been devoted to predict the usefulness of online discussions in a knowledge community. Such online discussions are quite different in nature from online reviews because an online knowledge community includes a rich set of social interactions that can be utilized for helpfulness prediction. In the following, we review relevant literature that uses text mining approaches to extract important features from online discussions and classify the usefulness of either threads or posts.

2.2. Computational Models of Information Usefulness

Besides conceptual models, IS researchers also study algorithmic solutions to evaluate helpfulness levels of online reviews. A representative example is Cao et al. (2011), who develop a text mining approach to extract semantic characteristics from review texts on CNET Download.com and subsequently examine the impacts of various features (basic, stylistic, and semantic) on the number of helpful votes online reviews receive. The authors demonstrate that semantic features are more important than other features when predicting the number of helpfulness votes. Other similar studies have explored the performance of other models such as text regression (Ngo-Ye and Sinha 2014), neural networks (Lee and Choeh 2014), and ensemble learning (Singh et al. 2017).

Information usefulness mining at the post level is closely related to another stream of research that assesses the quality of posts in online forums (Weimer et al. 2007, Shah and Pomerantz 2010). Many studies have considered it as an information retrieval task that ranks candidate answers by leveraging the relevance and relationship between a question and its answers (Cong et al. 2008, Surdeanu et al. 2008, Suryanto et al. 2009). There are also several studies that collect both textual and nontextual features of each post and use supervised learning algorithms to classify it into a binary group: correct answer or incorrect answer. For example, Hong and Davison (2009) and Wang et al. (2009) employ the support vector machine (SVM) algorithm to identify correct answers. They show that nontextual features rather than textual-based relevance features could help the classifier achieve a better performance. Additionally, a few studies have further incorporated the distinctive question-and-answer structure in online discussion threads to build more advanced learning algorithms to detect correct answers. Ding et al. (2008b) propose a general framework based on conditional random fields to find answers for questions from forum threads. Wang et al. (2010) use a deep belief network to model the semantic relevance of question–answer pairs using only word features and demonstrate good performance on finding the right answers.

The above-mentioned computational models are generally ad hoc; that is, most of the features in those studies are based on experience or intuitive understanding about what features are "important" (Wang and Strong 1996). One critical issue of that approach is the lack of a unified theoretical framework. We believe the design science paradigm in IS research (Walls et al. 1992, Hevner et al. 2004) can help guide a better design of information usefulness algorithm because it favors kernel theories to advise problem solving with a rigorous grounding and formulates a process to find the "optimal" solution design. This approach has been successfully applied to design a few information systems (Abbasi and Chen 2008a, De Sordi et al. 2016). Specifically, Abbasi and Chen (2008a) utilize a kernel theory of systemic functional linguistics and identify additional features (missed in previous studies) capable of presenting a rich array of information types for text analysis of computer-mediated communication. In our study, we closely follow the guidelines discussed in the design science research (Walls et al. 1992, Hevner et al. 2004) and develop a unified framework to classify the usefulness of information based on the behavioral theory of the knowledge adoption model. Therefore, both the conceptual and computational approaches of information usefulness are bridged together in our framework design. To meet the design research rigor, we perform a comprehensive performance evaluation of the system through an iterative search process and benchmark it with earlier studies to demonstrate the efficacy of the proposed framework.

3. Kernel Theory-Based Design: Usefulness Mining Using the Knowledge Adoption Model

The design science research advocates the application of rigorous methods to solve relevant problems (Hevner et al. 2004). Rigor is obtained from the use of the existing knowledge base of kernel theories and empirical observations. Such a knowledge base represents existing state-of-the-art theories, methods, and artifacts in the domain developed by experts.

Walls et al. (1992) propose an ISDT to formulate the integration of kernel theories into information system design. ISDT delineates four components in the design process, including kernel theories, meta-requirements, meta-design, and testable hypotheses. Kernel theories are natural or social theories governing the meta-requirements that need to be achieved by the product of design. The meta-design constructs a class of IT artifacts that are anticipated to meet the meta-requirements rigorously. Testable hypotheses are used to evaluate whether the proposed meta-design satisfies the meta-requirements. We outline our application of the ISDT in Table 2 and discuss the details of each of its four components in the following subsections.

3.1. Kernel Theory: Knowledge Adoption Model

There are several related theories that can be used to explain the usefulness of online communications. One prominent theory is the elaboration likelihood model (ELM), which provides a general framework for "organizing, categorizing, and understanding the basic processes underlying the effectiveness of persuasive communications" (Petty and Cacioppo 1986, p. 125). ELM states that two routes of information affect people's attitude toward persuasive information: the central route, which mainly deals with the content and topic of the message itself, and the peripheral route, which deals with the peripheral information related to the message sources and their characteristics (credibility, authority, commitment, etc.),

Table 2. Design Framework for an Information Usefulness Text Analytic System

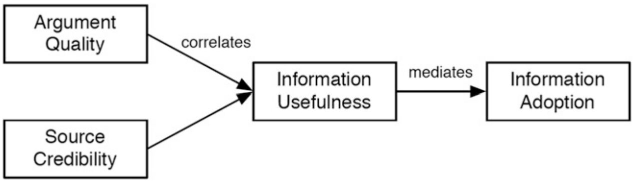
1. Kernel theory	KAM, which theorizes argument quality and source credibility as key determinants of information usefulness
2. Meta-requirements	Support for various computable dimensions that can represent the role of argument quality and source credibility in OKCs
3. Meta-design	• Construct a comprehensive set of computable metrics/features that are able to represent each dimension of argument quality and source credibility • Design an IT system (i.e., a text analytic framework that can extract features, model information usefulness, and finally train and test such a model)
4. Testable hypothesis	Evaluate the capabilities of our IT system. Specific testable hypotheses are as follows: • Hypothesis 1: An *IT system using KAM-based features can outperform the traditional text categorization approach for predicting information usefulness* • Hypothesis 2: *An IT system using KAM-based features can outperform previous studies, which used intuitive features, for predicting information usefulness*

as well as receiver moods, receivers' familiarity with the sources, etc. The central route is used when a receiver has the cognitive ability to process the message and contents. The peripheral route is used otherwise. Another similar and relevant theory is the heuristic-systematic model (HSM) of information processing (Chaiken 1980). In HSM, people may employ two models of information processing to process information: heuristic processing, where people use various heuristics (e.g., source credibility) to quickly reach a decision without much cognitive efforts, and systematic processing, where people involve comprehensive, analytic, and cognitive processes (related to the content and reliability of the messages) to make their judgments. As can be seen from the previous discussions, these two theoretical models share a lot of similarities. Both are dual models of information processing in understanding the persuasiveness of online communications.

Another theory is the knowledge adoption model (KAM), which models the perception of information usefulness as being affected by different elements of received information (Sussman and Siegal 2003). KAM extends the ELM to the context of electronic communication. The ELM model posits that a message can influence the recipient's attitude and behavior by its central and peripheral cues. Central cues refer to the arguments contained in the message, whereas peripheral cues refer to issues not directly related to the subject matter of the message. KAM considers message recipients' perception of information usefulness as the direct determinant of knowledge adoption. Determinants of perceived information usefulness for a given message include the perceived argument quality and the credibility of the source that posts the message (see Figure 2).

Either of the two constructs in KAM may independently lead to a positive or negative perception of usefulness for a particular message. Oftentimes, the two constructs jointly contribute to the knowledge adoption in a complex way. Sussman and Siegal's (2003) empirical analysis confirms the ELM theory that argument quality is a critical determinant of information usefulness when the message recipient is able to comprehend the message content well. Source credibility, however, becomes more important when the recipient is either unable or unwilling to process the message's content.

Figure 2. Model of Knowledge Adoption



Source. Adapted from Sussman and Siegal (2003, p. 52).

In this study, we chose the knowledge adoption model as a unified theory to identify key determinants of information usefulness. We view the voluntary tagging of useful solutions in an OKC as an information processing and knowledge adoption process where useful solutions are more likely to be adopted. It should be noted that KAM bears a lot of similarities with the ELM model and the HSM model. The key difference is that KAM postulates perceived usefulness as an important intermediate mediator variable that affects the adoption decision of a knowledge seeker. ELM and HSM, however, do not have the perceived usefulness as the mediator and lack the specificity to our problem context.

KAM theory has also been used as a theoretical basis in many other studies. These include information seeking and sharing behaviors in online investment communities (Park et al. 2014), information adoption from Wikipedia (Shen et al. 2013), usefulness of online customer reviews (Cheung et al. 2008), and its subsequent effect on purchase intentions (Erkan and Evans 2016). Zhu et al. (2014) and Liu and Park (2015) further develop metrics for a few dimensions of the KAM constructs (12 metrics in Zhu et al. and 15 metrics in Liu and Park), and they use those metrics in regression models to study the usefulness of online reviews. In the following, we propose a comprehensive set of quantitative metrics to measure each dimension of argument quality and source credibility in KAM.

3.2. Meta-Requirements: Argument Quality and Source Credibility Dimensions

Argument quality and source credibility are conceptual constructs that need to be operationalized as computational variables before we can utilize them in building our information usefulness analytic framework. To do that, we first research detailed dimensions from prior conceptual studies that can collectively represent argument quality and source credibility. Then, we select all the dimensions that are computationally feasible and use those as input variables of predict information usefulness.

Sussman and Siegal (2003) adopted only three instruments—*completeness*, *consistency*, and *accuracy*—from computer user satisfaction (Bailey and Pearson 1983, Doll and Torkzadeh 1991) to measure perceived argument quality. Those three dimensions belong to a broader construct of information quality in the literature. Wang and Strong (1996) consolidate existing studies about the quality of information content and proposed a holistic list of dimensions to assess information quality (IQ). The framework by Wang and Strong has been adopted to assist the development of computational models to measure helpfulness in online forums (Otterbacher 2009, Zhu et al. 2009). We follow suit and operationalize eight IQ dimensions to capture the quality of argument in a post: *appropriate amount of data*, *ease of understanding*, *relevancy*, *objectivity*, *timeliness*, *completeness*, *structure*, and *accuracy*.

Perceived source credibility refers to a message recipient's perception of the credibility and authority of the information source irrespective of the informational content (Chaiken 1980). Sussman and Siegal (2003) proposed two dimensions—*competence* and *trustworthiness*—as key factors of source credibility. The competence of a person can be measured through her *knowledge* (information and skills acquired through experience) and *expertise*. Given that knowledge in OKCs is intangible and often tacit (Ardichvili et al. 2006), we use *past experience* of community members as its proxy. *Expertise* is often considered as a more dominant dimension of source credibility (Wiener and Mowen 1986, Homer and Kahle 1990). In this study, we use the model of *expertise profiling* (Liu et al. 2012) to operationalize the expertise dimension of source credibility because it captures the domain specific expertise of a user. *Trustworthiness* refers to the believability of the source and is often considered as a key factor of source credibility (McGinnies and Ward 1980, Mowen et al. 1987). Studies on social capital (Tsai and Ghoshal 2008) have suggested that the social ties fostered among members through their online interactions can stimulate perceived trustworthiness. Hence, we use community network centrality measures such as in-degree and betweenness to measure trustworthiness and credibility of the source (Prell 2003).

3.3. Meta-Design: Argument Quality and Source Credibility Features

Whereas meta-requirements define the design goals (identify various dimensions of information usefulness), meta-design is to introduce a class of information technology (IT) artifacts that can achieve these goals (Walls et al. 1992). In the following, we propose a comprehensive set of metrics/features that collectively can evaluate the many dimensions of argument quality and source credibility discussed before.

3.3.1. Appropriate Amount of Data. Generally speaking, messages rich in content contain a lot of information. However, "too much" information can cause a readability issue and make it difficult for readers to find the information they want quickly. Therefore, an appropriate amount of data is necessary for high-quality information (Yang et al. 2005). We quantify textual information in a post using various lexical entries such as the number of characters, words, sentences, web links, math formulas, programming code, and so forth.

3.3.2. Ease of Understanding. The readability of online posts can be measured by characters-to-sentences and words-to-sentences ratios. Posts with high values on these measures are more difficult to comprehend (Otterbacher 2009). Additionally, we consider part-of-speech tagging and compute features such as the ratio of nouns, verbs, "wh"-type words, and punctuation to measure the syntactic complexity for understanding (Abbasi and Chen 2008b, Agichtein et al. 2008, Lu et al. 2010).

3.3.3. Relevancy. Online discussions often drift aimlessly from one topic to another (Potter 2009). The relevance of such discussions has been employed to predict the quality and helpfulness of online communities (Agichtein et al. 2009, Otterbacher 2009). We compute the semantic similarity between each response post and the original question or the other response posts in the thread using various approaches such as cosine similarity, Kullback–Leibler (KL) divergence, entropy, and perplexity to measure relevancy.

3.3.4. Objectivity. Information objectivity refers to the extent to which data are unbiased and impartial (Wang and Strong 1996). Knight and Burn (2005) and Lee et al. (2002) suggest that the subjective nature of information can influence its degree of quality. Previous studies (e.g., Pang and Lee (2004) and Ghose and Ipeirotis (2010)) have used sentiment and subjectivity analysis to identify and extract subjective signals from social media content. Following the approach in Ding et al. (2008a), we use a well-developed sentiment lexicon with about 6,800 words to create the sentiment features in our framework.

3.3.5. Timeliness. Timeliness is the extent to which the age of the data are appropriate for the task at hand (Wang and Strong 1996). It has been shown that the quality and helpfulness of online product reviews correlate with the time lapse of each post (Otterbacher 2009, Ghose and Ipeirotis 2010). To capture the timeliness dimension of a solution, we collect metrics such as the time of the post, the time lapse between a question and each response, and the time span between responses in a discussion thread.

3.3.6. Accuracy. Accuracy is an important measure in information systems success (Seddon 1997). In knowledge communities, the accuracy of data refers to the intrinsic quality of the textual content. Poor writings impair the readability of posts in OKCs and thus impact the perception of their usefulness. Previous studies have developed metrics to measure the accuracy of data in social media (Agichtein et al. 2008) and that of student writings such as punctuation and grammar mistakes (Shermis and Burstein 2003). We use LanguageTool (http://www.languagetool.org), an open source proofreading application programming interface, to identify various writing problems in text and compute our accuracy measure.

3.3.7. Structure. A post in a thread discussion can respond to any of its preceding posts. This is referred to as the discourse/reply structure of the thread. Aschoff et al. (2011) argue that the usefulness of an online forum can be measured by the discourse structure, which depends on the quantity and quality of the replies. Hong and Davison (2009) find that the position of a reply in the discourse is strongly related to its quality, and useful replies are normally not close to the bottom of a thread. Following Hong and Davison, we define a position metric for an answer post a and use it along with other discourse quality measures to capture the structure dimension of the usefulness of a post:

$$\text{Position}(a) = \frac{a\text{'s position to top of a thread}}{a\text{'s position to bottom of a thread}}.$$

3.3.8. Past Experience. According to the social capital theory, an individual develops cognitive capital as she interacts with other community members and gains hands-on experience over time (Wasko and Faraj 2005). As a result, an individual's past experience is an important factor for her reputation in knowledge communities. The metrics we use to measure past experience include an individual's tenure in the community, her past history of offering useful information, the ratio of questions and replies a user previously published in the community, and so forth.

3.3.9. Expertise. Expertise is considered as another type of cognitive capital. Individuals with a high level of expertise are more likely to provide useful advice in a virtual environment (Constant et al. 1996). We follow the expertise profiling approach (please see Liu et al. (2012) for details) to first obtain a question responder's expertise vector on a set of topics. Because the original question can also be projected onto the same set of topics, we can then compute the degree of expertise match between the responder and the question using similarity metrics (such as cosine similarity and KL divergence).

3.3.10. Trustworthiness. In addition to the social network structure measures (such as degree and betweenness centrality) previously used (Prell 2003, Tsai and Ghoshal 2008), we also incorporated closeness, cluster coefficient, PageRank, and HITS (Hyperlink-Induced Topic Search) hub and authority to measure the trustworthiness of the source. All of these measures are based on a social network of users in a knowledge community. We follow the approach in Gómez et al. (2008) to construct the social network where each community member corresponds to a node $u \in V$ and each directed edge $(u, v) \in E$ indicates the link from a replier to the original question seeker of the same thread. The network can be defined as a directed graph $G_s(V, E)$ with an adjacency matrix M, where $M(u, v) = n$ if person u has replied to person v in n distinct threads and $M(u, v) = 0$ otherwise. Ding et al. (2002) and Wassermann and Faust (1994) provide details of the methods to calculate these measures.

Table 3 summarizes the quantitative metrics for all the dimensions, inspired by the KAM theory, in our study. It should be noted that our approach of meta-design is generalizable to other types of OKCs because it makes no assumption on community structures and themes. However, to operationalize IT artifacts, the knowledge adoption process from contributors to question askers needs to be explicit and measurable. As such, communities with a focus on problem solving and troubleshooting such as Oracle Community and Apple Support Community are well suited for our model. By contrast, communities with a focus on brand building or idea sharing (see Table 1) are not applicable to this design because the underlining knowledge adoption process in these communities is implicit, and the metrics to model that process are also likely to be different.

3.4. Testable Hypotheses

As can be seen from earlier discussions and Table 3, we capture a much richer set of features and metrics, based on the KAM theory as well as a comprehensive review of relevant studies, than just lexical features used in traditional text categorization models. As such, we can hypothesize the following.

Hypothesis 1. *An IT system using KAM-based features can outperform the traditional text categorization approach for predicting information usefulness.*

For comparison purposes, we also summarize in Table 4 the dimensions used in prior computational models for either document classification or information retrieval. Most of those studies incorporate just a subset of the KAM dimensions and never consider timeliness and expertise. Additionally, the metrics used in prior models for each dimension are far less than those used in this study. Therefore, we can reasonably conjecture that our KAM-inspired system should outperform previous computational models of usefulness.

Table 3. Quantitative Metrics for Argument Quality and Source Credibility at Post Levels

Category	Dimensions	Metrics
Argument quality	F1: Appropriate amount of data (Abbasi and Chen 2008a, Otterbacher 2009)	Number of characters in a post Number of words in a post Number of unique words in a post Number of sentences in a post Number of nomenclature (e.g., programming code, math formula) in a post Number of web links in a post Number of quotations in a post
	F2: Ease of understanding (Otterbacher 2009)	Ratio of nouns, adjectives, comparatives, verbs, adverbs, punctuation, and symbols in a post Characters to sentences ratio in a post Words to sentences ratio in a post Number of “wh”-type words in a post Number of question marks in a post
	F3: Relevancy (Agichtein et al. 2009, Otterbacher 2009)	Cosine similarity between a question and one reply in a thread Query likelihood between a question and one reply in a thread KL divergence between a question and one reply in a thread Number of words overlapping between a question and one reply in a thread Entropy of one reply to all other replies in a thread Perplexity of one reply to all other replies in a thread Centroid of a reply to all other replies in a thread
	F4: Objectivity (Ding et al. 2008a)	Number of “thank” words of the OP (original poster) or other repliers to the replier in a post Ratio of positive and negative words of the OP to a replier in a thread Ratio of positive and negative words of repliers to other repliers in a thread Ratio of positive and negative words in a reply
	F5: Timeliness (Otterbacher 2009)	Time lapse between a question and each reply Time lapse between a reply and the previous post Reply post time as hour of day and day of week
	F6: Structure (Hong and Davison 2009, Aschoff et al. 2011)	A reply’s position metrics in terms of post order A reply’s position metrics in terms of replier order from top or bottom Number of responses to a reply from the OP Number of responses to a reply from other repliers If a poster is a replier or OP If a reply responds to the OP If a reply responds to other repliers If a reply is the first response to the OP If a reply is the second response to the OP
	F7: Accuracy (Agichtein et al. 2008)	Number and ratio of capitalization errors in a post Number and ratio of punctuation errors in a post Number and ratio of typos in a post Number and ratio of out-of-vocabulary words in a post Number and ratio of grammar errors in a post
Source credibility	F8: Past experience (Wasko and Faraj 2005)	Tenure (i.e., life time in a community) of the OP and repliers Number of questions and replies of a replier or OP previously published in the community Ratio of questions and replies of a replier or OP previously published in the community Number of replies a replier or OP published in a thread Repliers’ history of providing usefulness (no help, helpful, solved counts, and ratios) Authority metrics of a replier or OP based on his or her past publication in the community
	F9: Trustworthiness (Prell 2003, Tsai and Ghoshal 2008)	Social network analysis centrality measures of repliers (in-degree, out-degree, betweenness, closeness, cluster coefficient, PageRank, and HITS scores)
	F10: Expertise (Constant et al. 1996, Wasko and Faraj 2005)	Cosine similarity between OP or a replier’s expertise profile and the question KL divergence between OP or a replier’s expertise profile and the question OP or a replier’s LDA-based expertise score on the question Cosine similarity between a replier’s expertise profile and his or her reply KL divergence between a replier’s expertise profile and his or her reply A replier’s LDA-based expertise score on his or her reply

Table 4. Previous Computational Models (Classification (CL) or Information Retrieval (IR) Based) of Information Usefulness and Features Used

			Feature dimensions									
			AQ							SC		
Previous models	Level	Model	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Weimer et al. (2007)	Post	CL	√	√	√				√			
Hong and Davison (2009)	Post	CL						√				
Shah and Pomerantz (2010)	Post	CL	√					√		√		
Ding et al. (2008b)	Post	IR		√	√			√				
Cong et al. (2008)	Post	IR			√			√			√	
Surdeanu et al. (2008)	Post	IR			√							
Suryanto et al. (2009)	Post	IR			√							
Wang et al. (2009)	Post	IR	√		√	√						
Wang et al. (2010)	Post	IR			√							

Hypothesis 2. *An IT system using KAM-based features can outperform previous studies, which used intuitive features, for predicting information usefulness.*

4. Design Instantiation: A Usefulness Mining System

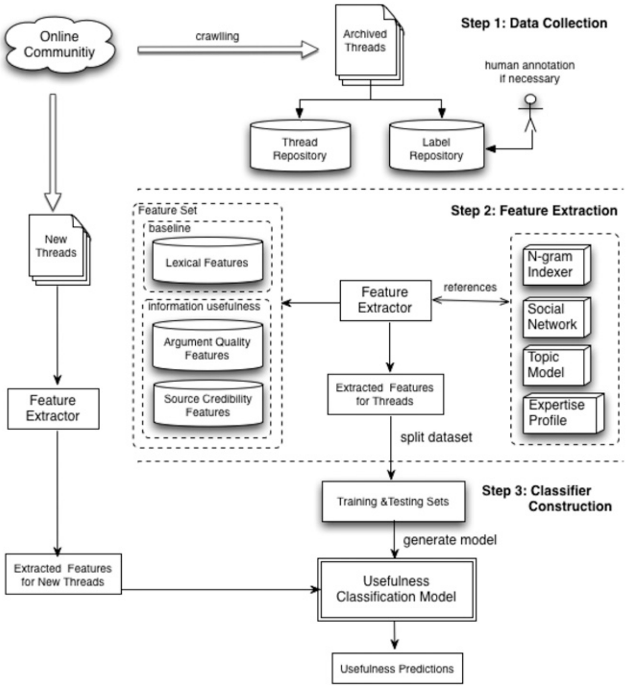
We now turn our focus to the design process itself: the architecture of the system. In our problem context, the new system is a text analytic framework to classify the usefulness of a solution in knowledge communities. Figure 3 details the architecture of our proposed framework, which includes data collection, feature extraction, and classifier construction.

Data collection deals with gathering necessary information on variables of interest. We developed a crawler to fetch all discussion threads from the targeted knowledge community. A parser was then employed to extract text messages, user information, and reply relationships of each discussion thread. These data were stored in a relational database; HTML tags and other metadata of web pages were ignored. Threads and posts with usefulness tags were identified as data instances for training usefulness classifiers.

Unstructured texts from the data collection module were subsequently converted to n-gram vectors to build the baseline lexical features. More complex feature extractors incorporating expertise profiles, topic models, and social networks were developed to extract various argument quality and source credibility features. It is worth noting that some of the features discussed earlier are easy to extract, whereas others may be computationally difficult and time consuming because they depend on the construction of social networks, topic models, or expertise profiles. That being said, the computation does not have to be in real time and can be done periodically offline. Finally, each post (or thread) i is converted into a tuple (x_i, y_i), where $x_i \in \mathfrak{R}^n$ is a vector of extracted features and $y_i \in L$ $(|L| > 1)$ is the usefulness label.

Four supervised learning algorithms, naïve Bayes, Decision Tree, Ada Boosting, and SVM, were chosen for our usefulness mining problem. These algorithms are representative of different types of classifiers and have been widely applied in various text mining tasks. Naïve Bayes is simple and fast yet often surprisingly effective. Decision Tree is frequently used because the tree structure of its learned model has good interpretability. We used the standard C4.5 implementation of Decision Tree. Ada Boosting (using Decision Stump as its weaker learners in this study) is one type of ensemble learning method, which has gained success in practice, including the famous Netflix competition. SVM is usually considered as a more robust algorithm. Specifically, the

Figure 3. System Architecture of a KAM-Based Usefulness Mining Framework in OKCs



sequential minimal optimization (SMO) implementation with polynomial kernels was used in our experiments. As is typical in any classifier learning process, we divided our experimental data into two subsets: a set of training data and a set of test data. A learning algorithm uses the training data to generate a classification (learned) model. The learned model is validated on the test data to assess their predictive power and performance.

5. Design Evaluation Settings

5.1. Evaluation Data

The utility, quality, and efficacy of our proposed framework need to be rigorously evaluated. To that end, we collected data from two representative online knowledge communities: Sun Forums (now part of Oracle forums, because Sun was acquired by Oracle) and Apple Discussions (now called Apple Support Community). They are the official OKCs for Oracle Corp. and Apple Inc., respectively. We selected the largest subforum from each community: the Java Programming subforum from Sun forums and the iPhone and Messaging forums from Apple Discussions. The discussions in iPhone and Messaging are product focused (iPhone) and consumer oriented. Anyone who has used iPhone can participate in the discussion. On the other hand, the Q&As in Java Programming require the understanding of the programming language, and this subforum is thus more professional oriented. Most of its members are IT professionals. We collected all threads and post messages up to June 6, 2009, for Java Programming and those up to April 1, 2010, for iPhone and Messaging. Table 5 reports the summary statistics of the two data sets.

To train the supervised learning algorithms in our framework, we need sufficient and reliable data with usefulness labels. Both OKCs allow community members to rate two levels of information usefulness: solved and helpful. It is fairly easy to parse and extract those labels from the crawled threads and posts. However, posts without any user ratings do not necessarily mean they are *not helpful* because of the possibility of "error of omission." To resolve this uncertainty, we asked domain knowledge experts to manually annotate information usefulness. To ensure the reliability of the manual annotation process, we first trained the domain experts using a few discussion threads and posts with user ratings and made sure the annotation of usefulness from different domain experts were consistent with the ratings from the original question askers. Additionally, we consider only those question-answering threads[3] with both "solved" and "helpful" posts. Presumably, question askers who initiated these threads perceived different levels of usefulness from received answers and were conscientious to provide reliable labels. Thus, it is reasonable to believe that the error of omission is minimized in this situation, and the workload of manual annotation can be eased. Table 6 describes the statistics of data sets for our post-level usefulness analysis.

5.2. Evaluation Setup

Given that a post can have different degrees of information usefulness (helpful or solved), we design two usefulness classification tasks. The first task is to predict all posts that are either helpful or solved to a question. The second task is to predict which post contains the best (i.e., solved) solution. These two classification tasks can be extended to analyze thread-level usefulness (see Online Appendices A and B).

We randomly sampled 600 threads (out of 6,672 for Apple Discussions and 2,666 for Oracle forums) with both solved and helpful posts as well as manually annotated not helpful posts to construct the post-level experiment data set. Positive instances here are solved and/or helpful solutions, whereas negative instances are those not helpful solutions. It is noted that the data (in terms of the number of positive versus negative instances) is highly imbalanced because the majority of the answers in a thread are usually not helpful. Previous literature demonstrated that a higher degree of data imbalance can lead to a higher error rate of classification (Japkowicz and Stephen 2002). To alleviate this issue, we employed the most commonly used oversampling strategy to replicate instances of minority class in the train data to make them balanced, whereas the test data maintain the original distribution to reflect the performance in real-world scenarios (Batista et al. 2004). There are also some other sampling strategies such as synthetic sampling, cost sensitivity learning, and active learning to reduce the data skewness from specific aspects (He and Garcia 2009). We do not investigate them because they are not the main focus of this study. For each post, 88 metrics of the 10 feature

Table 5. Summary Statistics of the Two Data Sets

Data set	No. of threads	No. of posts	No. of members	Average no. of replies per thread	Average no. of answerers (excluding OP) per thread	No. of words in data set
Apple	49,343	271,823	55,108	4.1306	3.1543	4,962,015
Oracle	70,488	440,708	36,687	4.1723	2.5137	7,374,557

Table 6. Characteristics of Post-Level Data Sets

Data attribute	Apple data	Oracle data
No. of questions (i.e., threads)	6,672	2,666
No. of answer posts	25,607	13,648
No. of "solved" posts	5,179 (20.2%)	1,703 (12.5%)
No. of "helpful" posts	3,536 (13.8%)	2,314 (17.0%)
No. of "not helpful" posts	16,892 (66.0%)	9,631 (70.5%)

dimensions (see Table 3) were computed. Those metrics become the independent variables in our classification tasks.

This setting allows us to conduct various experiments to systematically evaluate the efficacy of our proposed framework. We used precision, recall, and *f*-measure, which are standard information retrieval metrics (Manning et al. 2008), to mathematically evaluate the quality of trained classifiers. The *f*-measure, which accounts for both precision and recall, is a more popular metric to evaluate various classification models (Musicant et al. 2003). Another popular metric, classifier accuracy, was not used here because it can be easily biased by imbalanced data (Chawla et al. 2004). The definitions of the above metrics are given below; instances in the following equations are either threads or posts depending on the unit of analysis (thread level or post level) of information usefulness. A 10-fold cross-validation technique was employed to assess how a trained model will generalize to independent test data. All model evaluations presented in Section 6 use this technique:

$$precision = \frac{\text{No. of correctly predicted positive instances}}{\text{No. of predicted positive instances}},$$

$$recall = \frac{\text{No. of correctly predicted positive instances}}{\text{No. of actual positive instances}},$$

$$f\text{–}measure = \frac{2 \times precision \times recall}{(precison + recall)}.$$

5.3. Design Search

A good design is an iterative and incremental search process. From a computational perspective, it involves searching a design space with a sequence of operations to meet the final design objective(s). In our problem context, we seek to sequentially evaluate how each of the proposed feature sets will influence the model performance and whether a subset of those features can produce reasonably good design solutions that can be implemented in practice to predict information usefulness.

To that end, we first ran the traditional text categorization approach (baseline models), which relies only on lexical terms to construct feature sets (Yang and Pedersen 1997). This approach has worked well on categorizing a collection of documents into different topics (Sebastiani 2002, Blei et al. 2003). But we expect it would be inadequate because no single word can capture the semantics of information usefulness. Next we added each of the KAM feature dimensions to augment the traditional text classifier. Because each of those dimensions correlates with information usefulness according to the theory of KAM, we would expect improved model performance compared with the baseline models (Domingos 2012). Additionally, the metrics we proposed to measure argument quality and source credibility may be correlated, which can lead to model overfitting if all features are included in the model. Therefore we expect that a subset of selected features could produce a better model performance. Furthermore, given that deep learning (DL) has demonstrated salient advantages over traditional classifiers, we tested two DL models in our experiments. Last, but not the least, we conducted a series of experiments to benchmark and compare the performance of our model against prior studies.

6. Design Evaluation Results

6.1. Baseline Models

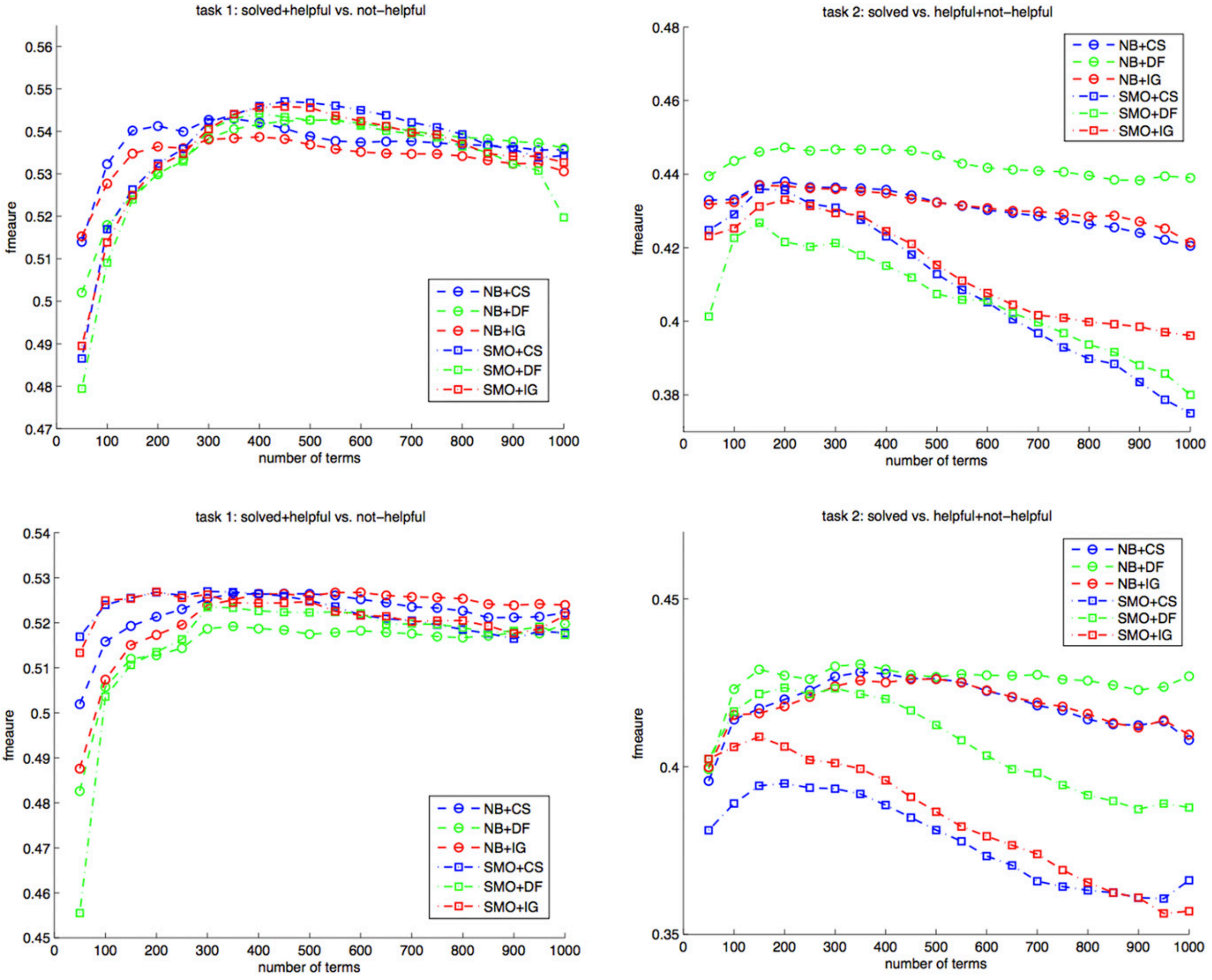
Following the method by Yang and Pedersen (1997), the feature space of each text consists of a vector of unique terms that appear in the text. The top *n* terms were selected based on the three most effective methods, χ^2-test (CS), document frequency (DF), and information gain (IG). Figure 4 presents the baseline model performance (*f*-measure) at the post level for two representative classification algorithms, naïve Bayes and SMO.[4] In each figure, the top and the bottom sections report the performance for the two classification tasks using the Apple Discussions data set and the Oracle forums data set, respectively.

As the graphs illustrate, all performance curves have a similar shape, with *f*-measure reaching its peak at about 150–450 terms and beginning to decline thereafter. We denote the set of selected terms that achieve the peak performance in the baseline models as our baseline feature set (F0). Figure 4 also demonstrates that the baseline models do not generate satisfactory results. The *f*-measures in the post-level classification were all below 0.6. When examining the most important (top 10) terms from feature selection using the baseline models, we noticed that those terms are not strong indicators of usefulness if considered only by their literal meanings. Overall, the traditional text categorization approach is not sufficient to accurately predict information usefulness from knowledge communities.

6.2. Baseline vs. KAM-Based Models

The advantage of KAM-based design is validated by testing its enhancement (in terms of the improvement

Figure 4. (Color online) Baseline Model Performance with Top *n* Terms for Apple (Top) and Oracle (Bottom) at the Post Level



in *f*-measure values) over the baseline models. We first built models (F0+Fx, $1 \leq x \leq 10$) by adding each KAM dimension (denoted as Fx, $1 \leq x \leq 10$) into the baseline model (F0) to verify that usefulness factors identified in the KAM theory can truly improve model performance. Then we examined the individual effect of the argument quality construct and the source credibility construct using two additional models (F0+AQ and F0+SC). Finally, a full model with all KAM features was assessed. The performances of these models are shown in Table 7 (Apple data set) and Table 8 (Oracle data set). The precision, recall, and *f*-measure of all four classification algorithms on the two classification tasks were reported in these tables. Our results demonstrate that all 10 feature dimensions based on KAM theory are important to improve the performance of predicting information usefulness in both data sets. Paired *t*-tests of *f*-measures on test folds of cross-validations, which are commonly used for information retrieval and machine learning model comparisons (Yang and Liu 1999) because of the test robustness (Hull 1993), show that each dimension alone can significantly improve model classification in at least one of the algorithms.

With the Apple data set, most argument quality dimensions (ease of understanding F2, relevancy F3, objectivity F4, timeliness F5, and structure F6) significantly improved *f*-measures over the baseline for all four classifiers and the two classification tasks. The dimension of an "appropriate amount of data" (F1) did not result in a significant improvement for task 1 except when using Ada Boosting. Adding an accuracy dimension (F7) has improved F0 alone, but its *t*-tests did not show significant differences. All source credibility dimensions (past experience F8, trustworthiness F9, and expertise F10) significantly improved *f*-measures over the baseline. When all feature dimensions were included, all classifiers except for naïve Bayes produced better classification performance than using only a single feature dimension. Ada Boosting

Table 7. Model Performance with KAM Features on the Apple Data Set at the Post Level

	Naive Bayes			C4.5			Ada Boosting			SMO		
	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.
Task 1: Helpful + Solved vs. Not helpful												
F0 (baseline)	0.570	0.503	0.531	0.520	0.560	0.538	0.506	0.535	0.520	0.565	0.516	0.539
F0+F1	0.520	0.600	0.557	0.531	0.565	0.547	0.444	0.924	**0.600*****	0.562	0.492	0.527
F0+F2	0.511	0.718	**0.597****	0.503	0.637	0.562	0.482	0.827	**0.609*****	0.566	0.538	**0.552***
F0+F3	0.558	0.777	**0.649*****	0.564	0.593	**0.578***	0.497	0.848	**0.627*****	0.608	0.686	**0.645*****
F0+F4	0.529	0.670	**0.591****	0.542	0.604	**0.571***	0.477	0.615	0.538	0.552	0.536	0.544
F0+F5	0.555	0.822	**0.663*****	0.617	0.756	**0.679*****	0.628	0.843	**0.720*****	0.606	0.834	**0.702*****
F0+F6	0.643	0.845	**0.730*****	0.636	0.819	**0.716*****	0.617	0.957	**0.751*****	0.571	0.724	**0.638*****
F0+F7	0.500	0.716	**0.587****	0.522	0.608	**0.561***	0.492	0.630	0.550	0.575	0.500	0.535
F0+F8	0.600	0.654	**0.626*****	0.576	0.743	**0.649*****	0.564	0.777	**0.654*****	0.581	0.655	**0.616*****
F0+F9	0.564	0.585	**0.574***	0.568	0.693	**0.624****	0.556	0.735	**0.633*****	0.545	0.721	**0.621*****
F0+F10	0.671	0.778	**0.721*****	0.665	0.787	**0.721*****	0.698	0.689	**0.693*****	0.681	0.733	**0.706*****
F0+AQ	0.607	0.837	**0.702*****	0.676	0.757	**0.707*****	0.699	0.791	**0.738*****	0.772	0.684	**0.713*****
F0+SC	0.602	0.748	**0.667*****	0.650	0.838	**0.732*****	0.624	0.962	**0.752*****	0.574	0.829	**0.679*****
All feature	0.645	0.775	**0.716*****	0.717	0.818	**0.761*****	0.756	0.867	**0.805*****	0.730	0.774	**0.745*****
Task 2: Solved vs. Helpful + Not helpful												
F0 (baseline)	0.434	0.457	0.445	0.377	0.462	0.415	0.377	0.508	0.433	0.436	0.452	0.444
F0+F1	0.410	0.660	**0.506****	0.447	0.453	**0.450***	0.375	0.837	**0.518****	0.481	0.348	0.404
F0+F2	0.455	0.612	**0.522****	0.440	0.488	**0.463***	0.433	0.732	**0.544*****	0.491	0.445	**0.467***
F0+F3	0.491	0.740	**0.590*****	0.494	0.572	**0.530****	0.457	0.765	**0.572*****	0.578	0.565	**0.571*****
F0+F4	0.402	0.658	**0.499***	0.391	0.603	**0.474****	0.402	0.588	0.478	0.392	0.477	0.430
F0+F5	0.495	0.780	**0.606*****	0.496	0.858	**0.628*****	0.495	0.885	**0.635*****	0.473	0.733	**0.575*****
F0+F6	0.514	0.825	**0.633*****	0.538	0.768	**0.633*****	0.518	0.982	**0.678*****	0.496	0.628	**0.554****
F0+F7	0.382	0.597	0.466	0.407	0.455	0.427	0.377	0.568	0.454	0.455	0.467	0.461
F0+F8	0.483	0.715	**0.577*****	0.486	0.797	**0.604*****	0.493	0.797	**0.609*****	0.523	0.725	**0.608*****
F0+F9	0.450	0.748	**0.562****	0.508	0.685	**0.583*****	0.478	0.797	**0.597*****	0.456	0.572	**0.507****
F0+F10	0.533	0.747	**0.622*****	0.580	0.677	**0.625*****	0.523	0.748	**0.616****	0.502	0.618	**0.554****
F0+AQ	0.493	0.788	**0.607*****	0.504	0.808	**0.626*****	0.534	0.924	**0.677*****	0.532	0.685	**0.599*****
F0+SC	0.482	0.865	**0.619*****	0.514	0.810	**0.629*****	0.475	0.937	**0.630*****	0.443	0.757	**0.559****
All features	0.506	0.803	**0.621*****	0.599	0.722	**0.655*****	0.603	0.827	**0.697*****	0.511	0.753	**0.608****

Notes. Significant results compared with the baseline model are highlighted in bold. Prec., precision; Rec., recall; *f*-Mea., *f*-measure.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

achieved the best performance with *f*-measures for tasks 1 and 2 reaching 0.805 and 0.697, respectively. Analysis using the Oracle data set yielded similar observations (see Table 8). All in all, our rigorous analysis of various classification algorithms using two separate data sets provided solid evidence and strong support of argument quality and source credibility in determining the information usefulness levels.

We further replicated the same evaluation process on the thread-level usefulness classification. The results of those analyses are presented in Tables B.4 and B.5 of Online Appendix B. Our extensive evaluations confirm Hypothesis 1: the IT system using KAM-based features outperforms the traditional text categorization approach for predicting information usefulness.

An interesting observation from post-level analysis indicates that the source credibility dimensions (F8, F9, and F10), on average, were more effective than the argument quality dimensions (F1–F7). A few of them, such as appropriate amount of data, ease of understanding, relevancy, and objectivity, occasionally did not result in a significant improvement in model performance. The comparisons between all argument quality features (F0+AQ) and all source credibility features (F0+SC) at the post level showed similar outcomes. This indicated that the source of the answer is a more important determinant of usefulness than its content. In the context of traditional media, data volume often positively correlates with data quality (Wang and Strong 1996). The same, however, does not always apply in OKCs. Table 9 shows two sample discussion threads that exemplify how (a) short messages can sometimes be more useful than longer ones, and (b) answers receiving acknowledgement (e.g., positive sentiments) do not necessarily imply that the answer is helpful.[5]

6.3. Feature Selection and Optimal Subset of Features

Feature selection (determining an "optimal" set of features), along with feature engineering (determining

Table 8. Model Performance with KAM Features on the Oracle Data Set at the Post Level

	Naive Bayes			C4.5			Ada Boosting			SMO		
	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.
Task 1: Solved + Helpful vs. Not helpful												
F0 (baseline)	0.541	0.502	0.521	0.492	0.505	0.498	0.495	0.479	0.487	0.524	0.526	0.525
F0+F1	0.505	0.592	**0.545***	0.504	0.601	**0.548****	0.473	0.745	**0.579*****	0.548	0.464	0.502
F0+F2	0.531	0.520	0.526	0.490	0.481	0.485	0.490	0.638	**0.554****	0.625	0.319	0.423
F0+F3	0.490	0.617	**0.546***	0.486	0.514	0.500	0.475	0.785	**0.592*****	0.521	0.580	0.549
F0+F4	0.450	0.780	**0.571****	0.400	0.786	**0.530***	0.426	0.720	**0.536****	0.470	0.673	0.553
F0+F5	0.505	0.807	**0.621*****	0.621	0.648	**0.634*****	0.613	0.734	**0.668*****	0.591	0.832	**0.691*****
F0+F6	0.567	0.875	**0.688*****	0.587	0.782	**0.671*****	0.563	0.907	**0.667*****	0.567	0.758	**0.649*****
F0+F7	0.544	0.533	0.538	0.455	0.561	0.498	0.488	0.539	0.507	0.536	0.511	0.521
F0+F8	0.553	0.840	**0.667*****	0.558	0.747	**0.639*****	0.527	0.933	**0.674*****	0.564	0.791	**0.658*****
F0+F9	0.557	0.893	**0.685*****	0.549	0.767	**0.640*****	0.506	0.978	**0.667*****	0.547	0.786	**0.645*****
F0+F10	0.637	0.752	**0.690*****	0.601	0.699	**0.646*****	0.583	0.730	**0.648*****	0.630	0.749	**0.684*****
F0+AQ	0.622	0.737	**0.666*****	0.609	0.702	**0.645*****	0.654	0.740	**0.686*****	0.627	0.738	**0.672*****
F0+SC	0.561	0.809	**0.661*****	0.583	0.751	**0.650*****	0.564	0.997	**0.720*****	0.577	0.842	**0.682*****
All features	0.625	0.757	**0.693*****	0.611	0.788	**0.688*****	0.669	0.787	**0.716*****	0.620	0.828	**0.709*****
Task 2: Solved vs. Helpful + Not helpful												
F0 (baseline)	0.408	0.445	0.426	0.364	0.435	0.396	0.383	0.328	0.354	0.394	0.423	0.408
F0+F1	0.412	0.668	**0.510****	0.449	0.585	**0.508*****	0.418	0.640	**0.506*****	0.415	0.413	0.414
F0+F2	0.433	0.535	**0.478***	0.366	0.378	0.372	0.358	0.483	**0.411***	0.472	0.242	0.320
F0+F3	0.405	0.695	**0.512*****	0.397	0.568	**0.467***	0.414	0.575	**0.482*****	0.445	0.565	**0.498****
F0+F4	0.323	0.740	0.450	0.290	0.682	0.407	0.277	0.573	0.374	0.304	0.582	0.400
F0+F5	0.406	0.707	**0.516*****	0.432	0.585	**0.497***	0.440	0.822	**0.573*****	0.433	0.765	**0.553*****
F0+F6	0.437	0.807	**0.567*****	0.459	0.802	**0.579*****	0.438	0.828	**0.573*****	0.443	0.720	**0.548*****
F0+F7	0.355	0.727	**0.476***	0.404	0.417	0.410	0.359	0.657	**0.452****	0.442	0.420	**0.431***
F0+F8	0.431	0.795	**0.559*****	0.417	0.883	**0.567*****	0.415	0.922	**0.572*****	0.438	0.713	**0.543*****
F0+F9	0.421	0.840	**0.561*****	0.414	0.710	**0.523*****	0.408	0.890	**0.560*****	0.411	0.675	**0.511*****
F0+F10	0.557	0.708	**0.575*****	0.494	0.645	**0.559*****	0.472	0.773	**0.582*****	0.534	0.590	**0.561*****
F0+AQ	0.445	0.802	**0.573*****	0.513	0.670	**0.581*****	0.502	0.735	**0.597*****	0.489	0.695	**0.574*****
F0+SC	0.508	0.752	**0.606*****	0.525	0.607	**0.563*****	0.517	0.757	**0.614*****	0.518	0.618	**0.563*****
All features	0.496	0.734	**0.592*****	0.531	0.703	**0.605*****	0.529	0.763	**0.624*****	0.526	0.704	**0.602*****

Notes. Significant results are highlighted in bold. Prec., precision; Rec., recall; *f*-Mea., *f*-measure.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

what features to use), is as important as the prediction ability of the final classifier. We herein performed feature selection analysis to find the "optimal" subset of features that achieve the best model performance based on *f*-measure. As pointed out by Chandrashekar and Sahin (2014), finding the globally optimal feature subset from 2^n possible subsets is an NP-hard problem and needs to run for an almost infinite amount of time when the number of features n is large. For this reason, many feature selection approaches have been proposed to find a "locally best feature subset" within reasonable time. Our purpose here is not to solve this NP-hard problem but to show that feature selection, which finds a "locally optimal subset," can further improve the performance of usefulness prediction, and thus, feature selection is a necessary step in algorithm design. We adopted a filter-type feature selection algorithm that ranks all 338 post-level features (250 lexical and 88 KAM features) by information gain and then filters out irrelevant features from low rank in a stepwise manner. Through n filtering steps and trials, this process can find a subset that is "optimal" for this selection method. This feature selection approach is common for text mining problems because it is simple and efficient on a data set with a large number of features (Yang and Pedersen 1997). It should be noted that different classification algorithms likely select a different optimal subset of features. Similarly, different features are likely to be chosen for different problem contexts as a result of the idiosyncratic characteristics of a knowledge community.

Tables 10 and 11 report the number of selected features and the performance of the best model we obtained in the selection process. In Online Appendix C, we listed the top 20 features ranked by information gain to demonstrate the most important features for each model. It can be seen that feature selection indeed resulted in improved performance (*f*-measure) and at the same time drastically reduced the number of features. Naïve Bayes and SMO classifiers with feature selection produced significant performance

Table 9. Sample Discussion Threads from the Apple Discussions Community

Posts	Usefulness
Panel A: Thread 1 (https://discussions.apple.com/thread/1018581, accessed August 2009)	
Q: So I was sucked into the hype.. now whats a good screen protector?	
A1: shieldzone.com. the best. I promise.	Helpful
A2: I have protector from shieldzone.com and they are great! On the other hand, it might not need screen protection: http://www.pcworld.com/video/id,545-page,1-bid,0/video.html?tk=synd_macworld.	Solved
A3: I bought the crystal film today at the apple store, it stays in place way better than you would imagine with no sticky stuff, and it dosent make the touch screen any harder to use, in my opinion it made the touch screen work BETTER because the plastic is easier to slide on than glass. I also bought the incase rubber case for it, and I am using the two together so it is fully protected.	Not helpful
Panel B: Thread 2 (https://discussions.apple.com/thread/1017348, accessed August 2009)	
Q: I just setup my new iphone and everything is great except that there is no sound at all on the external speakers... I don't hear the ringtone, speaker phone or anything for that matter. The headphones do work though. I have tried all the settings and tried using the manual as well... Am I missing something or should I just take it back...???	
A1: Try to remove the plastic if you haven't already, a lot of people were having sound problems with the plastic on the front.	Not helpful
Q: didn't work... Thanks though...	

improvements. Improvements of C4.5 and Ada Boosting, however, are not always salient and consistent. We also observed similar feature selection results in thread-level analysis (see Table B.6 and B.7 in Online Appendix B).

6.4. KAM-Based Model vs. Deep Learning

The recent rise of deep learning methods has changed the landscape of natural language processing (LeCun et al. 2015). In addition to its high accuracy, DL has the potential to process raw data and automatically discover intricate structures from the data, which traditionally requires considerable feature engineering efforts and domain expertise. Therefore, a natural question to ask is how does our theory-driven design algorithm pit against popular deep learning models.

We implemented two DL models using the Python Keras library. The first model is the long short-term memory (LSTM) recurrent neural network (Sundermeyer et al. 2015), which can capture temporal patterns automatically from plain text. We test it to see whether this deep learning technique's automatic feature engineering capability can outperform the theory-guided manual feature design; if so, then theory-based usefulness model design may not be even necessary. Our LSTM network was structured with an embedding layer, with 200 dimensions mapping each word to an embedding; an LSTM layer (also dropout layers preceding and succeeding), with 100 neurons well suited to process sequence data such as text; and finally, a dense output layer using a single neuron to generate a 0 or 1 output. The second DL model is the standard convolutional neural network (CNN) (Krizhevsky et al. 2012), which is more appropriate for both textual and nontextual features. We constructed the network with a layer that convolves input feature space into a tensor output, a flatten layer, and a dense output layer to generate predicted classes.

Table 12 presents the performance comparison of the two DL models with the baseline model and the best KAM-based model (using Ada Boosting). The comparison shows that both DL models consistently performed better than the baseline model, but our

Table 10. Model Performance with Feature Selections for the Apple Data Set at the Post Level

	Naive Bayes			C4.5			Ada Boosting			SMO		
	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.
Task 1: Helpful + Solved vs. Not helpful												
Best subset (no. of features)	0.638	0.970	**0.770**** (3)	0.726	0.890	**0.796*** (18)	0.742	0.908	0.817 (67)	0.728	0.815	0.769 (42)
All	0.645	0.775	0.716	0.717	0.818	0.761	0.756	0.867	0.805	0.730	0.774	0.745
Task 2: Solved vs. Helpful + Not helpful												
Best subset (no. of features)	0.503	0.953	**0.658*** (207)	0.602	0.815	**0.687*** (46)	0.625	0.822	0.710 (34)	0.538	0.766	**0.632*** (107)
All	0.506	0.803	0.621	0.599	0.722	0.655	0.603	0.827	0.697	0.511	0.753	0.608

Notes. Pairwise *t*-tests were performed on *f*-measures using selected features versus all features. Significant results are highlighted in bold. Prec., precision; Rec., recall; *f*-Mea., *f*-measure.

*$p < 0.05$; **$p < 0.01$.

Table 11. Model Performance with Feature Selections for the Oracle Data Set at the Post Level

	Naïve Bayes			C4.5			Ada Boosting			SMO		
	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.	Prec.	Rec.	*f*-Mea.
						Task 1: Helpful + Solved vs. Not helpful						
Best subset (no. of features)	0.576	1.000	**0.731*** (2)	0.576	1.000	**0.731*** (2)	0.673	0.802	0.732 (80)	0.622	0.885	0.729 (12)
All	0.625	0.757	0.693	0.611	0.788	0.688	0.669	0.787	0.716	0.620	0.828	0.709
						Task 2: Solved vs. Helpful + Not helpful						
Best subset (no. of features)	0.503	0.749	0.602 (8)	0.443	1.000	0.614 (3)	0.545	0.804	0.649 (61)	0.539	0.769	**0.633*** (112)
All	0.496	0.734	0.592	0.531	0.703	0.605	0.529	0.763	0.624	0.526	0.704	0.602

Notes. Pairwise *t*-tests were performed on *f*-measures using selected features versus all features. Significant results are highlighted in bold. Prec., precision; Rec., recall; *f*-Mea., *f*-measure.
*$p < 0.05$.

best KAM-based model still outperformed them. This result is not a total surprise. Even though LSTM is able to capture semantics, syntax, and even sequence information from text (as opposed to only a bag of words in the baseline model), it fails to extract more theoretically meaningful features related to information usefulness. When deep learning (i.e., CNN) works directly on all extracted features, its performance improved over LSTM but was still lower than the KAM-based model. This could be because the number of learnable parameters in the multilayer CNN is too large for our experimental data sets. To function well, DL models usually require millions of observations, which exceeds the total available usefulness labels in a typical OKC. We obtained similar results in our thread-level analysis (Table B.8 in Online Appendix B), except that LSTM performed better than CNN in some cases.

6.5. KAM-Based Model vs. Previous Studies

Here, we compare the performance of our KAM-based model against existing algorithms that seek to identify high-quality posts or correct answers from various online communities. We reproduced three influential approaches in the literature: Weimer et al. (2007), Hong and Davison (2009), and Shah and Pomerantz (2010). We evaluated the performance of these three algorithms on our data sets. For fair comparison, we ran their classifiers along with the four classifiers in this study using the features proposed in their studies and picked only the best model in each approach.

Table 13 presents the performance comparisons of our model against the three approaches. It is clear that the KAM-based framework beat all three state-of-the-art algorithms with significant margins on both data sets. It should be pointed out that Hong and Davison (2009) show that they were able to produce near-perfect (over 90% accuracy) results on their data sets, but their algorithm performed poorly in our experiments. To figure out the reasons behind this, we examined the values of authorship versus position in our Apple and Oracle data sets to understand how these two features correlate to different levels of usefulness. In Hong and Davison, most of the correct answers

Table 12. Benchmarking of KAM-Based Models with DL Models at the Post Level

	Apple			Oracle		
Studies	Precision	Recall	*f*-Measure	Precision	Recall	*f*-Measure
			Task 1: Helpful + Solved vs. Not helpful			
Baseline 1 (plain text)	0.565	0.516	0.539	0.434	0.457	0.445
Baseline 2: LSTM (plain text)	0.566	0.829	0.673	0.538	0.818	0.649
Baseline 3: CNN (all features)	0.711	0.782	0.745	0.655	0.806	0.722
KAM based (all features)	0.756	0.867	**0.805**	0.669	0.787	0.716
KAM based (selected features)	0.742	0.908	**0.817**	0.673	0.802	**0.732**
			Task 2: Solved vs. Helpful + Not helpful			
Baseline 1 (plain text)	0.524	0.526	0.525	0.408	0.445	0.426
Baseline 2: LSTM (plain text)	0.430	0.729	0.541	0.406	0.712	0.518
Baseline 3: CNN (all features)	0.592	0.780	0.670	0.499	0.705	0.581
KAM based (all features)	0.603	0.827	**0.697**	0.529	0.763	**0.624**
KAM based (selected features)	0.625	0.822	**0.710**	0.545	0.804	**0.649**

Note. Significant results compared with the baseline model are highlighted in bold.

Table 13. Benchmarking of KAM-Based Models with Previous Studies at the Post Level

	Apple			Oracle		
Studies	Precision	Recall	*f*-Measure	Precision	Recall	*f*-Measure
Task 1: Helpful + Solved vs. Not helpful						
Weimer et al. (2007)	0.523	0.710	0.602	0.475	0.719	0.572
Hong and Davison (2009)	0.574	**0.896**	0.699	0.567	0.825	0.672
Shah and Pomerantz (2010)	0.631	0.725	0.675	0.611	0.644	0.627
KAM based	**0.746****	0.871	**0.804****	**0.619**	**0.884***	**0.728***
Task 2: Solved vs. Helpful + Not helpful						
Weimer et al. (2007)	0.382	0.618	0.472	0.323	0.703	0.442
Hong and Davison (2009)	0.447	**0.883**	0.594	0.443	0.787	0.566
Shah and Pomerantz (2010)	0.469	0.747	0.576	0.414	0.660	0.509
KAM based	**0.613****	0.832	**0.706****	**0.525***	**0.830***	**0.643****

Note. Significant results compared with the baseline model are highlighted in bold.
$^{*}p < 0.05$; $^{**}p < 0.01$.

are close to the top positions (i.e., earliest answers), and the authors of these answers are usually senior community members who prefer to write replies rather than asking a question (Hong and Davison 2009). However, we did not observe this pattern in our data. In fact, we found that the earliest answers in Apple Discussions and Oracle forums exhibited various usefulness levels. Overall, we can confirm Hypothesis 2 on the superiority of our KAM-based model over previous approaches using limited and intuitive features.

7. Discussions and Conclusions

Many users are now turning to OKCs to search useful information and knowledge for their needs. As the knowledge base increases, it becomes more challenging to sift through the tremendous amount of unstructured data to identify useful information and solutions in online knowledge communities. Hence more effective knowledge management tools are necessary to fulfill the informational needs of the consumers. In this paper, we proposed a theory-driven text analytic framework that can automatically extract important features and predict multiple levels of information usefulness in OKCs. Using data from two of the largest problem-solving communities, Apple Discussions and Oracle forums, we conducted a series of rigorous and robust tests to evaluate the performance of our model. The results demonstrated superior efficacy of our designed artifact.

7.1. Implications for Research

Our text analytic framework introduces a comprehensive set of features based on the theory of knowledge adoption model. Previous studies have identified a few important features and built learning algorithms based on those features to predict correct or high-quality answers in online forums. But none of them developed their models based on a holistic approach for a multiple-level analysis of information usefulness. Our framework provides a solid foundation for future studies that aim to develop applications of distilling useful knowledge from online communities.

Guided by ISDT, this research showcases a design process that utilizes the conceptual KAM theory to specify the meta-requirements of representing various dimensions of information usefulness and construct the meta-design of computational models to implement each dimension. Whereas kernel theories drawn from natural or social sciences should govern the design process itself (Walls et al. 1992), how to integrate kernel theories into design science research needs to be further explored. Specifically, how to operationalize conceptual theories into building computational models in IT artifacts is a challenge and may vary case by case. Our research illustrates a detailed approach of applying the KAM theory to the design of a text analytics system for information usefulness. We discussed how computational features could be derived from conceptual dimensions to represent information usefulness based on the KAM theory, and we demonstrated that our theory-driven design can outperform other intuition-driven designs significantly. We hope this paper can shed light on future studies of theory-driven design in IS research.

This study is inspired by the design science philosophy, which advocates the application of rigorous methods and theoretical foundations in both the construction and evaluation of the designed artifact. It is, to the best of our knowledge, among the first to apply the well-established KAM theory in the process of designing an analytic framework that has advanced existing state-of-the-art approaches significantly. To that end, our study showcased an approach combining multiple disciplines (behavioral, empirical, design science, and technical) to provide new insights into an important and relevant business problem. Furthermore, our results demonstrate that even powerful deep learning algorithms cannot learn all latent features from raw data automatically. Kernel theories such as KAM can play critical roles in understanding a complicated process such as the perception of usefulness and thus can help to guide a better algorithm design to automate this process.

Our results also offer some interesting feedback to the KAM theory. Argument quality (AQ) and source credibility (SC) seem to play different roles in usefulness prediction at the thread level (AQ is more important) and the post level (SC is more effective). One possible reason is that arguments in a thread discussion can proceed multiple rounds to result in a solution. For the post level, a post in the early stage of argument can have good argument quality but is not considered as useful. A later post can solve the problem but is very concise (i.e., the argument quality is not high). Thus, we posit that *argument progress* may be a new moderator of information usefulness overlooked by existing theories. Furthermore, the selected features reveal that the "structure" (F6) and "past experience" (F10) dimensions are most helpful to post usefulness prediction. The metrics for these two dimensions are introduced in this paper to capture the social interactions among community members. The constructs in KAM (Sussman and Siegal 2003) were originally derived from email communications, which lack a social engagement component. In this regard, we urge future researchers to revisit both argument quality and source credibility, and consider adding a social interaction construct to KAM.

Several future extensions are possible. The features adopted in this study are built on the existing knowledge of conceptual and computational models for information usefulness. Unique features with appropriate metrics could further improve the validity and applicability of our framework. For example, this study does not include the "consistency" dimension of argument quality. Consistency refers to the extent to which data are always presented in the same format and are compatible with previous data (Wang and Strong 1996). This dimension is not measurable in OKC because question-and-answer revisions are not publicly available. It is also worth examining the correlations among the feature dimensions in-depth. A factor analysis might be helpful in further developing an optimized set of feature dimensions. Additionally, performance evaluation analysis showed that a feature dimension might behave differently in various classification algorithms. It would be interesting to evaluate the sensitivity of feature dimensions with respect to each classification algorithm. Last, but not least, recent advances in DL techniques have significantly improved the performance of algorithms. Yet the interpretability of those algorithm output is still low. We believe that studies that combine advanced ML algorithms and extant IS theories (information quality theory, elaboration likelihood model, information processing theory, pervasion theory, etc.) to solve relevant business problems with good explainability to business decision makers have a bright future. For instance, our algorithm design referenced theories involving information quality and source credibility, both of which could be important concepts when developing explainable ML algorithms to predict fake news on social media platform, user adoption of physician replies in online healthcare communities, and crowdfunding project success using project descriptions. We encourage future IS researchers to pursue more research along this line.

7.2. Implications for Practice

A practical implication of the study is for sponsors of various OKCs to rethink their knowledge management strategy and build a system that can better serve users' knowledge needs. Many knowledge communities have resorted to the wisdom of the crowd concept to seek a user's collective opinions on the usefulness of a solution. Some even dedicated a significant amount of human resources to monitor and manually organize a knowledge repository of the community. As discussed earlier, both approaches have their limitations. An alternative approach to tackle this important problem is to use machine learning techniques to discover useful knowledge automatically. Our proposed framework was rigorously evaluated using data from two popular communities and demonstrated superior capability. It provides practical guidance on how to build such a machine learning-based system.

The framework described here can also be used to improve users' experience of searching for useful knowledge. OKCs usually provide a search functionality, which allows users to type in some keywords and retrieve a list of threads that are ranked by their *relevance* to query keywords. However, relevant information does not always mean it is useful to fulfill users' knowledge needs. We propose that future search functionalities should rank search results by not only *relevance* but also *usefulness*. Additionally, search engines in OKCs can incorporate multiple levels of usefulness into the search results. When returning useful threads, the system can be designed to present the most useful posts as snippets under each thread.

Our framework can be applied to other domains as well. Chatbots that conduct machine generated conversations with humans have the potential to transform the experience of customer services (Accenture 2016). A direct application of our framework would be to construct a set of useful answers from companies' OKCs to bootstrap the development of the customer-facing chatbots. Google envisions its next-generation search engine as a tool to help people find any information at any time via question answering (Macrae 2014). This requires the system to understand any human-language question that users input and to return the right answer immediately. One challenge is that there may be many candidate answers for a question and thus how to decide the right set of answers for the question. This study offers an alternative lens through which we can build a robust system to detect useful answers.

Acknowledgments

The authors thank the entire review team for their constructive feedback; this paper has benefited tremendously from those suggestions.

Endnotes

[1] An exception is StackOverflow, which was reported to have almost 70% of its questions marked as "solved" (Anderson et al. 2012). One reason is that StackOverflow has a well-designed reputation system (see Anderson et al. (2012) for more details) and massive human mediations to encourage its members to provide usefulness assessments. Unfortunately, most OKCs do not have such sophisticated functionalities and affluent resources.

[2] From a conceptual perspective, post-level usefulness should determine thread-level usefulness because it is a reflection as to whether a thread contains useful or solved posts. From a computational and theoretical perspective, however, they can be represented by different models because a thread-level model can aggregate data from multiple posts to obtain different features and thus may produce better results than just using data from a single post.

[3] It is noteworthy that not all threads are about question answering. Classifiers have been proposed to filter out nonquestion threads (Cong et al. 2008, Hong and Davison 2009).

[4] Decision Tree and Ada Boosting were also experimented and produced similar results. For brevity, we did not report them on the plots.

[5] In OKCs, we frequently observed that users tend to acknowledge others' replies even though they were not really helpful.

References

Abbasi A, Chen H (2008a) CyberGate: A design framework and system for text analysis of computer-mediated communication. *MIS Quart.* 32(4):811–837.

Abbasi A, Chen H (2008b) Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inform. Systems* 26(2):1–29.

Accenture (2016) Chatbots in customer service. Accessed June 7, 2018, https://goo.gl/1W2ah9.

Agichtein E, Liu Y, Bian J (2009) Modeling information-seeker satisfaction in community question answering. *ACM Trans. Knowledge Discovery Data* 3(2):10:1–10:27.

Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. Najork M, ed. *Proc. Internat. Conf. Web Search Web Data Mining* (ACM, New York), 183–194.

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Discovering value from community activity on focused question answering sites. Yang Q, ed. *Proc. 18th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 850–858.

Ardichvili A, Maurer M, Li W, Wentling T, Stuedemann R (2006) Cultural influences on knowledge sharing through online communities of practice. *J. Knowledge Management* 10(1):94–107.

Aschoff F-R, Schaer V, Schwabe G (2011) Where should I send my post? The concept of discourse quality in online forums and its dependency on membership size. Kjeldskov J, Paay J, eds. *Proc. 5th Internat. Conf. Communities Tech.* (ACM, New York), 69–78.

Baek H, Ahn J, Choi Y (2012) Helpfulness of online consumer reviews: Readers' objectives and review cues. *Internat. J. Electronic Commerce* 17(2):99–126.

Bailey JE, Pearson SW (1983) Development of a tool for measuring and analyzing computer user satisfaction. *Management Sci.* 29(5): 530–545.

Batista GEAPA, Prati RC, Monard MC (2004) A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations* 6(1):20–29.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.

Cao Q, Duan W, Gan Q (2011) Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems* 50(2):511–521.

Chaiken S (1980) Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Personality Soc. Psych.* 39(5):752–766.

Chandrashekar G, Sahin F (2014) A survey on feature selection methods. *Comput. Electronic Engrg.* 40(1):16–28.

Chawla NV, Japkowicz N, Kotcz A (2004) Special issue on learning from imbalanced data sets. *SIGKDD Explorations* 6(1):1–6.

Cheung CM, Lee MK, Rabjohn N (2008) The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet Res.* 18(3):229–247.

Cheung CMK, Lee MKO, Lee ZWY (2013) Understanding the continuance intention of knowledge sharing in online communities of practice through the post-knowledge-sharing evaluation processes. *J. Amer. Soc. Inform. Sci. Tech.* 64(7): 1357–1374.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Chua AYK, Banerjee S (2015) Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *J. Assoc. Inform. Sci. Tech.* 66(2):354–362.

Cong G, Wang L, Lin C-Y, Song Y-I, Sun Y (2008) Finding question-answer pairs from online forums. Myaeng S-H, Oard DW, Sebastiani F, Chua T-S, Leong MK, eds. *Proc. 31st Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 467–474.

Constant D, Sproull L, Kiesler S (1996) The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organ. Sci.* 7(2):119–135.

Cox A (2005) What are communities of practice? A comparative review of four seminal works. *J. Inform. Sci.* 31(6):527–540.

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* 13(3): 319–340.

De Sordi JO, Meireles M, Luiz de Oliveira O (2016) The text matrix as a tool to increase the cohesion of extensive texts. *J. Assoc. Inform. Sci. Tech.* 67(4):900–914.

de Vries S, Kommers P (2004) Online knowledge communities: Future trends and research issues. *Internat. J. Web Based Comm.* 1(1):115–123.

Ding X, Liu B, Yu PS (2008a) A holistic lexicon-based approach to opinion mining. Proc. Internat. Conf. Web Search Web Data Mining (ACM, New York), 231–240.

Ding S, Cong G, Lin C-Y, Zhu X (2008b) Using conditional random fields to extract contexts and answers of questions from online forums. *Proc. 46th Annual Meeting Assoc. Comput. Linguistics: Human Language Tech.* (Association for Computational Linguistics, Stroudsburg, PA), 710–718.

Ding C, He X, Husbands P, Zha H, Simon HD (2002) PageRank, HITS and a unified framework for link analysis. Jarvelin K, Beaulieu M, Baeza-Yates R, Myaeng SH, eds. *Proc. 25th Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 353–354.

Doll WJ, Torkzadeh G (1991) The measurement of end-user computing satisfaction: Theoretical and methodological issues. *MIS Quart.* 15(1):5–10.

Domingos P (2012) A few useful things to know about machine learning. *Comm. ACM* 55(10):78–87.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* 45(4):1007–1016.

Erkan I, Evans C (2016) The influence of eWOM in social media on consumers' purchase intentions: An extended approach to information adoption. *Comput. Human Behav.* 61(August):47–55.

Fan W, Wallace L, Rich S, Zhang Z (2006) Tapping the power of text mining. *Comm. ACM* 49(9):76–82.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* 19(3):291–313.

Ghose A, Ipeirotis PG (2010) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Gómez V, Kaltenbrunner A, López V (2008) Statistical analysis of the social network and discussion threads in Slashdot. Huai J, Chen R, Hon H, Liu Y, Ma W-Y, Tomkins AS, Zhang X, eds. *Proc. 17th Internat. Conf. World Wide Web* (ACM, New York), 645–654.

Gray PH (2001) A problem-solving perspective on knowledge management practices. *Decision Support Systems* 31(1):87–102.

Gregor S (2006) The nature of theory in information systems. *MIS Quart.* 30(3):611–642.

Haas MR, Criscuolo P, George G (2015) Which problems to solve? Online knowledge sharing and attention allocation in organizations. *Acad. Management J.* 58(3):680–711.

He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Trans. Knowledge Data Engrg.* 21(9):1263–1284.

Hearst MA (1999) Untangling text data mining. Dale RF, Church KW, eds. *Proc. 37th Annual Meeting Assoc. Comput. Linguistics Comput. Linguistics* (Association for Computational Linguistics, Morristown, NJ), 3–10.

Hevner AR, March ST, Park J, Ram S (2004) Design science in information systems research. *MIS Quart.* 28(1):75–105.

Homer PM, Kahle LR (1990) Source expertise, time of source identification, and involvement in persuasion: An elaborative processing perspective. *J. Advertising* 19(1):30–39.

Hong L, Davison BD (2009) A classification-based approach to question answering in discussion boards. Sanderson M, Zhai CX, Zobel J, Allan J, Aslam JA, eds. *Proc. 32nd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 171–178.

Hong H, Xu D, Wang GA, Fan W (2017) Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems* 102(October):1–11.

Hu N, Liu L, Zhang JJ (2008) Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Inform. Tech. Management* 9(3):201–214.

Huang P, Tafti A, Mithas S (2018) Platform sponsor investments and user contributions in knowledge communities: The role of knowledge seeding. *MIS Quart.* 42(1):213–240.

Huang AH, Chen K, Yen DC, Tran TP (2015) A study of factors that contribute to online review helpfulness. *Comput. Human Behav.* 48(July):17–27.

Hull D (1993) Using statistical testing in the evaluation of retrieval experiments. Korfhage R, Rasmussen E, Willett P, eds. *Proc. 16th Annual Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 329–338.

Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. *Intelligent Data Anal.* 6(5):429–449.

Karimi S, Wang F (2017) Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems* 96(April):39–48.

Knight S, Burn JM (2005) Developing a framework for assessing information quality on the World Wide Web. *Informing Sci.* 8(5): 159–172.

Kraaijenbrink J, Wijnhoven F, Groen A (2007) Towards a kernel theory of external knowledge integration for high-tech firms: Exploring a failed theory test. *Tech. Forecasting Soc. Change* 74(8): 1215–1233.

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates, Red Hook, NY), 1097–1105.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.

Lee S, Choeh JY (2014) Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems Appl.* 41(6):3041–3046.

Lee GK, Cole RE (2003) From a firm-based to a community-based model of knowledge creation: The case of the Linux kernel development. *Organ. Sci.* 14(6):633–649.

Lee YW, Strong DM, Kahn BK, Wang RY (2002) AIMQ: A methodology for information quality assessment. *Inform. Management* 40(2):133–146.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Inform. Systems Res.* 19(4):456–474.

Liu Z, Park S (2015) What makes a useful online review? Implication for travel product websites. *Tourism Management* 47(April): 140–151.

Liu X, Wang GA, Johri A, Zhou M, Fan W (2012) Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Inform. Systems Frontiers* 16(4):715–727.

Lu Y, Tsaparas P, Ntoulas A, Polanyi L (2010) Exploiting social context for review quality prediction. Rappa MA, Jones P, Freire J, Chakrabarti S, eds. *Proc. 19th Internat. Conf. World Wide Web* (ACM, New York), 691–700.

Macrae P (2014) Google going where no search engine has gone before: Amit Singhal. *Phys.org* (November 5), http://phys.org/news/2014-11-google-amit-singhal.html.

Majchrzak A, Wagner C, Yates D (2013) The impact of shaping on knowledge reuse for organizational improvement with Wikis. *MIS Quart.* 37(2):455–470.

Malik MSI, Hussain A (2018) An analysis of review content and reviewer variables that contribute to review helpfulness. *Inform. Processing Management* 54(1):88–104.

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, UK).

McGinnies E, Ward CD (1980) Better liked than right. *Personality Soc. Psych. Bull.* 6(3):467–472.

Mowen JC, Wien JL, Joag S (1987) An information integration analysis of how trust and expertise combine to influence source credibility and persuasion. *Adv. Consumer Res.* 14(1):564.

Mudambi SM, Schuff D (2010) What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quart.* 34(1):185–200.

Musicant DR, Kumar V, Ozgur A (2003) Optimizing F-measure with support vector machines. Russell I, Haller S, eds. *Proc. 16th Internat. Florida Artificial Intelligence Res. Soc. Conf.* (American Association for Artificial Intelligence, Menlo Park, CA), 356–360.

Ngo-Ye TL, Sinha AP (2014) The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems* 61(May):47–58.

Otterbacher J (2009) 'Helpfulness' in online communities: A measure of message quality. Olsen DR, Arthur RB, eds. *Proc. 27th Internat. Conf. Human Factors Comput. Systems* (ACM, New York), 955–964.

Pan Y, Zhang JQ (2011) Born unequal: A study of the helpfulness of user-generated product reviews. *J. Retailing* 87(4):598–612.

Pang B, Lee L (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Scott D, ed. *Proc. 42nd Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 271–278.

Park JH, Gu B, Man Leung AC, Konana P (2014) An investigation of information sharing and seeking behaviors in online investment communities. *Comput. Human Behav.* 31(February):1–12.

Petty RE, Cacioppo JT (1986) The elaboration likelihood model of persuasion. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (Springer, New York), 1–24.

Potter A (2009) Constructive chaos: Topic management in asynchronous learning networks. *Internat. J. Tech. Knowledge Soc.* 5(3):1–12.

Prell CL (2003) Community networking and social capital: Early investigations. *J. Comput.-Mediated Comm.* 8(3), https://doi.org/10.1111/j.1083-6101.2003.tb00214.x.

Racherla P, Friske W (2012) Perceived "usefulness" of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Res. Appl.* 11(6):548–559.

Sarker S, Valacich JS (2010) An alternative to methodological individualism: A non-reductionist approach to studying technology adoption by groups. *MIS Quart.* 34(4):779–808.

Sebastiani F (2002) Machine learning in automated text categorization. *ACM Comput. Surveys* 34(1):1–47.

Seddon PB (1997) A respecification and extension of the DeLone and McLean model of IS success. *Inform. Systems Res.* 8(3):240–253.

Shah C, Pomerantz J (2010) Evaluating and predicting answer quality in community QA. Crestani F, Marchand-Maillet S, Chen H, Efthimiadis EN, Savoy J, eds. *Proc. 33rd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 411–418.

Shen X-L, Cheung CMK, Lee MKO (2013) What leads students to adopt information from Wikipedia? An empirical investigation into the role of trust and information usefulness. *British J. Educational Tech.* 44(3):502–517.

Shermis MD, Burstein JC, eds. (2003) *Automated Essay Scoring: A Cross-Disciplinary Perspective* (Lawrence Erlbaum Associates, Mahway, NJ).

Singh JP, Irani S, Rana NP, Dwivedi YK, Saumya S, Roy PK (2017) Predicting the "helpfulness" of online consumer reviews. *J. Bus. Res.* 70(January):346–355.

Sundermeyer M, Ney H, Schluter R (2015) From feedforward to recurrent LSTM neural networks for language modeling. *IEEE/ACM Trans. Audio Speech Language Processing* 23(3):517–529.

Surdeanu M, Ciaramita M, Zaragoza H (2008) Learning to rank answers on large online QA collections. Arisoy E, Maier W, Inoue K, eds. *Proc. 46th Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 719–727.

Suryanto MA, Lim EP, Sun A, Chiang RHL (2009) Quality-aware collaborative question answering: Methods and evaluation. Baeza-Yates R, Boldi P, Ribeiro-Neto B, Cambazoglu BB, eds. *Proc. Second ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 142–151.

Sussman SW, Siegal WS (2003) Informational influence in organizations: An integrated approach to knowledge adoption. *Inform. Systems Res.* 14(1):47–65.

Tsai W, Ghoshal S (2008) Social capital and value creation: The role of intrafirm networks. *Acad. Management J.* 41(4):464–476.

Walls JG, Widmeyer GR, El Sawy OA (1992) Building an information system design theory for vigilant EIS. *Inform. Systems Res.* 3(1):36–59.

Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. *J. Management Inform. Systems* 12(4):5–33.

Wang B, Liu B, Sun C, Wang X, Sun L (2009) Extracting Chinese question-answer pairs from online forums. *Proc. 2009 IEEE Internat. Conf. Systems, Man Cybernetics* (IEEE, Piscataway, NJ), 1159–1164.

Wang B, Wang X, Sun C, Liu B, Sun L (2010) Modeling semantic relevance for question-answer pairs in web social communities. Hajič J, Carberry S, Nivre J, eds. *Proc. 48th Annual Meeting Assoc. Comput. Linguistics* (Association for Computational Linguistics, Stroudsburg, PA), 1230–1238.

Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quart.* 29(1):35–57.

Wassermann S, Faust K (1994) *Social Network Analysis: Methods and Applications* (Cambridge University Press, New York).

Weimer M, Gurevych I, Mühlhäuser M (2007) Automatically assessing the post quality in online discussions on software. Ananiadou S, ed. *Proc. 45th Annual Meeting ACL Interactive Poster Demonstration Sessions* (Association for Computational Linguistics, Stroudsburg, PA), 125–128.

Wenger E (1998) *Communities of Practice: Learning, Meaning, and Identity* (Cambridge University Press, Cambridge, UK).

Wiener JL, Mowen JC (1986) Source credibility: On the independent effects of trust and expertise. *Adv. Consumer Res.* 13(1):306–310.

Yang Y, Liu X (1999) A re-examination of text categorization methods. Gey FC, Hearst MA, Tong R, eds. *Proc. 22nd Internat. ACM SIGIR Conf. Res. Development Inform. Retrieval* (ACM, New York), 42–49.

Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. Fisher DH, ed. *Proc. 14th Internat. Conf. Machine Learn.* (Morgan Kaufmann Publishers, San Francisco), 412–420.

Yang Z, Cai S, Zhou Z, Zhou N (2005) Development and validation of an instrument to measure user perceived service quality of information presenting web portals. *Inform. Management* 42(4):575–589.

Yin D, Bond S, Zhang H (2014) Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quart.* 38(2):539–560.

Zhou S, Guo B (2017) The order effect on online review helpfulness: A social influence perspective. *Decision Support Systems* 93(January):77–87.

Zhu Z, Bernhard D, Gurevych I (2009) A multi-dimensional model for assessing the quality of answers in social Q&A sites. Technical Report TUD-CS-2009-0158, Technische Universität Darmstadt, Darmstadt, Germany.

Zhu L, Yin G, He W (2014) Is this opinion leader's review useful? Peripheral cues for online review helpfulness. *J. Electronic Commerce Res.* 15(4):267–280.